Frequent homozygous deletions of the CDKN2A locus in somatic cancer tissues: analysis of chromosomal aberrations in cell cycle regulator genes

Research Thesis

Presented in partial fulfillment of the requirements for graduation
with research distinction in the undergraduate colleges of The Ohio State University

by

Abdulaziz Hamid

The Ohio State University
November 2018

Project Advisor: Dr. Ruben Petreaca, Department of Molecular Genetics

ABSTRACT

Homozygous deletions (HD) of human CDKN2A and neighboring regions on the p arm of Chromosome 9 have been previously reported in some cancers but a pan-cancer analysis of the aberrations of this locus is lacking. Here we analyzed the copy number variations that include CDKN2A locus using data acquired with an Affymetrix SNP6.0 array and deposited on the Catalogue of Somatic Mutations in Cancer (COSMIC) database. We find that inactivation of CDKN2A by HD is not cancer specific. A majority of HDs of this locus have a median range of 1,255,650 base pairs. We then mapped the positions of breakpoints of these deletions on both the telomere and centromere proximal sides of CDKN2A. Remarkably, most of the telomere proximal breakpoints map to a narrow region of the chromosome where the genes MTAP and MIR31HG are located. The centromere proximal breakpoints of the deletions are distributed over a wider chromosomal region. This comprehensive analysis shows that inactivation of CDKN2A by homozygous deletion is not cancer specific but rather determined by the chromosomal region.

ACKNOWLEDGEMENTS

I would like to thank my research advisor Dr. Ruben Petreaca for his tremendous help and support during my journey in conducting independent research in Molecular Genetics. I have been conducting research since Spring semester of 2017 until Autumn semester of 2018. I am extremely grateful for having the opportunity to have worked with a humble, intelligent and caring professor. During my research experience with Dr. Petreaca I have gained valuable research skills I will continue to use in my future academic career.

I am also thankful to Dr. Ryan Yoder for serving on my committee as a co-advisor.

I would like to thank my family, especially my parents (Umi Amin and Bazi Amin) for their on-going support and love during my college career. I would not be able to succeed without them.

I would also like to thank Dr. John Maharry, the director of the Honors Committee and other members of the committee for their generous financial support.

LIST OF FIGURES

TABLE OF CONTENTS

INTRODUCTION

Deoxyribonucleic acid (DNA) is a molecule composed of two antiparallel strands of sugar phosphate groups that twist to form a double helix held together by hydrogen bonds. Hydrogen bonds are between complementary nitrogenous base pairs, adenine (A) to thymine (T) and cytosine (C) to guanine (G). Two hydrogen bonds connect A to T and three hydrogen bonds connect C to G. DNA is the genetic material that serves as instructions for all cellular processes. The information encoded in DNA (the genome) is transcribed into mRNA and translated into proteins. During cellular replication a copy of the genome is passed onto progeny (TUBBS AND NUSSENZWEIG 2017). DNA is constantly exposed to exogenous and endogenous damage. Exogenous agents include ionizing radiation (IR), X-rays, ultraviolet (UV) light, oxidative damage during metabolism, base hydrolysis and other various chemical agents (FRIEDBERG 2008). Endogenous or spontaneous damage arises mainly from mistakes made by the DNA replicative machinery when the genome is copied during cell division. Exogenous and endogenous agents can cause genetic changes in the DNA genome that can lead to accumulation of mutations and cancer (FRIEDBERG 2008). The DNA damage and repair mechanisms play a critical part in maintaining genome stability (WANG AND VASQUEZ 2017).

Types of DNA damage and their consequences

The DNA genome is subjected to many forms of DNA damage daily (**Figure 1**) (Biology Learning Center). Indeed, in a single day approximately 70,000 lesions occur in each human cell, and a clear majority (75%) of the lesions are DNA single stranded breaks (SSBs) (LINDAHL AND BARNES 2000; FRIEDBERG 2008). A single stranded DNA break occurs when one strand that makes up the double stranded DNA is severed. In a few instances, SSBs may be converted into DNA double stranded breaks (DSBs). A double stranded DNA break arises when both strands of the DNA helix are severed. Both SSBs and DSBs can result from exogenous agents such as ionizing radiation (IR) induced directly or indirectly from reactive oxygen species (ROS). Both types of breaks compromise the genomic integrity of the genome. DSBs are the most dangerous type of lesions if left unrepaired or repaired incorrectly (WARD 1988). Unlike SSBs which can be repaired from the non-broken template, DSBs must be repaired by looking for homology elsewhere in the genome. Therefore, repair of DSBs is more mutagenic than SSBs. When a DSB is repaired incorrectly, it can induce mutations and chromosomal rearrangements such as deletions, duplications, inversions and translocations. A deletion or duplication occurs when a segment of a chromosome is lost or copied respectively. An inversion occurs when a segment of a chromosome is reversed, and the orientation is flipped. Translocation occurs when one segment of a chromosome is exchanged with a segment from another chromosome. These events can lead to genomic instability which may lead to cell death or cancer (JACKSON AND BARTEK 2009). Although most DSBs are repaired correctly and do not lead to chromosomal aberrations, incorrect repair can occur when the cell encounters mutations in components of the repair machinery.

DSBs can be categorized as one-ended or two-ended DSBs. Exogenous damage such as reactive oxygen species (ROS) generated by ionization radiation results in two-ended DSBs. Two-ended DSBs are generated when both strands of the DNA double helix are broken simultaneously at the same position or near the same proximity. Endogenous damage to DNA usually results in one-ended DSBs. A vast majority of one-ended

DSBs occurs during DNA replication, when a single stranded DNA break left unrepaired encounters a DNA replication fork triggering fork stalling or collapse. One-ended DSBs can also arise when a DNA replication fork stalls and a nuclease degrades one of the chromosome arms. Unlike two ended-DSBs, one ended DSBs are more problematic to repair because the second end does not exist (RANJHA *et al.* 2018).

The processing and repair of DSBs becomes very important because unrepaired or inappropriately repaired DNA DSBs can lead to mutations and chromosomal aberrations that may result in cell death or ultimately give rise to cancer. DNA repair pathways are highly conserved in eukaryotic organisms. DNA damage and repair were first studied in model organisms such as in the yeasts (*Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*) and further extrapolated to higher level eukaryotic organisms such as humans. All eukaryotic organisms share similar DNA repair mechanisms but depending on the complexity of the organism, mechanism of break formation and repair as well as the preferred stage of the cell cycle may differ (PASTINK *et al.* 2001). This introduction will focus on DNA double stranded breaks and the different mechanisms of repair in human cells with emphasis on how inappropriate repair of chromosomal breaks leads to alteration of chromosomal copy number and cancer. These forms of chromosomal aberrations or copy number variations (CNVs) vary in size from 1 kilobases to several megabases in length. A high level of CNVs is associated with cancer.

Cancer is a complex disease characterized by the lifelong accumulation of mutations and inappropriately repaired DNA damage resulting in alteration of the cellular machinery and leading to uncontrolled cell division. In addition to mutation accumulation (one base pair change), cancer cells show a high level of chromosomal aberrations such as loss of heterozygosity, homozygous deletion, and amplification. Loss of heterozygosity is a genetic event that leads to loss of one copy of a genetic region. Homozygous deletion constitutes the loss of both copies of the genetic region. Amplification is an event that results in gain of more than two copies of a genetic region. The chromosome copy number in normal diploid cells is fine-tuned at two per cell. Any alteration from this ratio will result in inappropriate gene expression of the type seen in cancer cells.

In all cancer cells, part of the genome is either amplified or deleted to some degree (**Figure 2**), a process known as cellular transformation. There is a complex pattern between amplification and deletion that occurs across samples of cancer cells. Certain regions of the genome are more prone to have many chromosomal aberrations while other regions only have a few. Some regions of the genome consist of large deletions compared to short deletions. And some regions of the genome comprise large amplifications compared to other regions that show only a few amplifications. There may be a correlation between these observations (KASHIWAGI AND UCHIDA 2000). Not every gene that is amplified or deleted causes genomic disorders. Rather, mutations of certain key genes serve as drivers for rapid mutation accumulation. Duplication and deletions involving these genes gives rise to cancer and other genomic disorders (LUPSKI 1998; KENNEDY *et al.* 2012).

Cell cycle regulation and defects in cancer cells

Normal cells grow, divide and commit programmed cell death which is controlled by the cell cycle. The cell cycle must be tightly regulated as it governs DNA replication and cell

division. The cell cycle is made up of two distinct phases, interphase (G_1, S and G_2 phases) and the mitotic (M) phase. The M phase can be further divided into prophase, metaphase, anaphase, telophase and cytokinesis. During interphase, specifically in S phase DNA replication occurs and during the mitotic (M) phase the replicated DNA is divided equally between the two daughter cells (PINES 1994a; PINES AND HUNTER 1994).

Throughout the cell cycle there are cell cycle checkpoints in place to maintain genomic integrity. Cells are in constant threat of endogenous and exogenous damage. Most damage is detected during or immediately after S-phase therefore the most important checkpoint is G2 and to a second degree an intra-S phase checkpoint. The checkpoints serve as a surveillance system, they asses the cell progression and detect DNA damage. If a checkpoint detects DNA damage, it can do one of two things: arrest cell cycle and recruit machineries to repair the damage or, if damage cannot be fixed stimulate apoptosis which is programmed cell death. Cancerous cells are not responsive to many of these checkpoints (VISCONTI *et al.* 2016). Mutations in two types of cell cycle regulators, proto-oncogenes and tumor suppressor genes may give rise to cancer. Proto-oncogenes are genes that are responsible for cell growth. A mutation in a proto-oncogene can activate it to become an oncogene that results in rapid cell growth. Tumor suppressor genes (or checkpoint genes) are genes responsible for slowing down cell division, allowing damaged DNA to be repaired or to allow for cell death if the damage cannot be repaired. A mutation in a tumor suppressor gene can allow bypass of cellular surveillance machinery.

In addition to checkpoints, the cell cycle is further regulated by certain cyclin dependent kinases (CDKs). CDKs belongs to a family of serine/threonine kinases, that relies on the binding of specific regulatory subunits known as cyclins for kinase activity (PINES 1994b). The CDK phosphorylates and modifies the function of different targets throughout the cell cycle to drive the cell cycle forward. Therefore, the CDK can be thought as the master regulator of the cell cycle. Mutations in CDKs or their associated cyclins have been identified in cancer cells. Only certain CDK-cyclin complexes are found to be directly involved in regulating the cell cycle. Specific cyclins are synthesized and destroyed in the cell at certain stages of the cell cycle to allow passage of the cell through the different stages of the cycle (MALUMBRES AND BARBACID 2009; MALUMBRES *et al.* 2009). Different CDK-cyclin complexes are involved in initiating DNA synthesis (S phase) and mitotic (M) phase (PINES 1994a).

CDKs can also bind inhibitor proteins which function to inhibit cyclin dependent kinase activity. For example, certain CDK inhibitors prevents the transition from G1 to S and G2 to M in the presence of DNA damage. CDK inhibitors are involved in cell cycle arrest, to allow cells to repair damage DNA and prevent damage DNA being passed down to the daughter cells(MALUMBRES AND BARBACID 2009; MALUMBRES *et al.* 2009). Therefore, they can be thought as checkpoint genes. The CDKN2A (cyclin dependent kinase inhibitor 2A) is a CDK1 inhibitor that inhibits the G1/S transition.

Cell cycle deregulation is one of the hallmarks of cancer. Tumor cells accumulate mutations, that results in three cell cycle defects: unscheduled proliferation, genomic instability and chromosomal instability. Genomic instability (GIN) refers to the increase tendency of mutations in the genome and chromosomal aberrations. Chromosomal instability (CIN), a type

of genomic instability, is characterized by altered chromosomal number (MALUMBRES AND BARBACID 2009; MALUMBRES *et al.* 2009). Mis-regulation of CDKs have been associated with unscheduled proliferation, GIN and CIN (MALUMBRES AND BARBACID 2005).

Mechanisms of DNA damage repair

To maintain genomic stability, higher level eukaryotic organisms have developed multiple mechanism to repair damaged DNA with minimal loss of genetic information (JACKSON AND BARTEK 2009; LUKAS AND BARTEK 2009). The three major DNA excision repair pathways that repair base alterations are base excision repair (BER), mismatch repair (MMR), and nucleotide excision repair (NER). BER is responsible for the repair of altered DNA bases. MMR is responsible for recognizing and repairing base mismatches during DNA replication. NER is responsible for removing and repairing DNA damages induced by ultraviolet light (UV), consisting of thymidine dimers. DNA excision repair pathways maintain the genomic integrity of DNA, prevent mutations and development of cancer and genomic disorders.

The mechanisms for repair of DNA DSBs differ from the DNA excision repair pathways. Higher level eukaryotic organisms are equipped with two major processes for repair of DNA double stranded breaks, homologous recombination (HR) and nonhomologous DNA end joining (NHEJ) (TUBBS AND NUSSENZWEIG 2017). In homologous recombination (HR) the missing information resulting from a DSB is copied from the undamaged sister chromatid (**Figure 3**). Thus, HR provides high fidelity, ensuring accurate repair and restoring genetic information. Since HR is template-dependent, HR operates during the S and G2 phase after DNA replication when the sister chromatid is available (ORTHWEIN *et al.* 2015). HR can repair both one- and- two ended DSBs (KOWALCZYKOWSKI 2015; CHANG *et al.* 2017). Homologous recombination is used to rescue stalled or collapsed replication forks which generally produce one ended breaks that cannot be repaired by the NHEJ pathway which requires two ended breaks (RANJHA *et al.* 2018). We and others have previously shown that mutations in key genes that control the fidelity of replication and facilitate accurate DNA damage repair increase the numbers of chromosomal aberrations or CNVs (LI *et al.* 2013; BHATTACHARJEE AND NANDI 2016; HROMAS *et al.* 2016).

The end-joining pathways can be further classified as canonical non-homologous end-joining and alternative nonhomologous end-joining (alt-NHEJ) also known as microhomology-mediated end-joining (MMEJ). NHEJ involves the direct ligation of the DSB ends and requires no homology for repair. NHEJ is template-independent, it does not require the sister chromatid for repair and can operate at any stage of the cell cycle. NHEJ is subjected to large deletions and chromosomal rearrangements leading to genetic information loss (CHANG *et al.* 2017). When both ends of a single DSB are present, they can simply be ligated together via NHEJ, however when multiple DSB breaks are present this becomes problematic for NHEJ. NHEJ can rejoin wrong DNA ends together leading to chromosomal rearrangements such as translocation (KOWALCZYKOWSKI 2015). The microhomology-mediated DSB repair pathway have similarities with both NHEJ and HR. MMEJ also involves the direct ligation of the DSB ends but requires 2-5 base pairs (bp) of homology for repair. The repair pathways by these mechanisms are fast but potentially mutagenic (RANJHA *et al.* 2018).

Whether the cell employs HR or NHEJ for repair of DNA DSBs depends largely on the nature of the DSB ends. Extensive DNA end resection commits the cell to repair via HR and at

the same time inhibits NHEJ. NHEJ is not able to ligate DNA ends in the presence of DNA end resection. HR utilizes the exposed tracts of ssDNA for homology search (SYMINGTON AND GAUTIER 2011; CHAPMAN *et al.* 2012; SHIBATA 2017). Activation of CDK-cyclin complex initiates DNA end resection in the S and G2 phase of the cell cycle (WANG *et al.* 2018).

Inactivation of the tumor suppressor CDKN2A in cancer cells

As noted earlier, it has been long appreciated that the most represented mutations in cancer cells are of genes that control the cell cycle and those involved in DNA damage repair. For example, p53 mutations are identified in over 50% of cancers (BIEGING *et al.* 2014). The second most represented genetic aberration in cancers is the CDKN2A locus which is found on Ch9p21 (ORTEGA *et al.* 2002; KIM AND SHARPLESS 2006; POI *et al.* 2015). CDKN2A encodes three different tumor suppressors: ARF which promotes p53 stabilization, and p15 and p16 which are inhibitors of Cyclin D/CDK4/6. The types of chromosomal aberrations observed at this locus in cancer cells fall in three general categories, simple mutations (amino acid substitution, insertion or deletion), promoter methylation, and deletions larger than 100 base pairs (STRANSKY *et al.* 2011; MOUNTZIOS *et al.* 2014; CICENAS *et al.* 2017). Remarkably, in some cancers, over half of the chromosomal aberrations that span the CDKN2A locus are large deletions (CONRAD *et al.* 2010; GAST *et al.* 2010; LEE *et al.* 2015).

Recently CDKN2A breakpoints have been reported in melanoma (XIE *et al.* 2016) and pancreatic cancers (NORRIS *et al.* 2015a), but we still lack a comprehensive analysis in different cancers. The COSMIC (Catalogue of Somatic Mutations in Cancer) database deposits cancer genomes data from different sources, including the ICGC (International Cancer Genome Consortium). In this report, we queried the COSMIC database which reports CNVs generated using an Affymetrix SNP6.0 array and analyzed with two algorithms, PICNIC and ASCAT and generated a map of the breakpoints in the CDKN2A deletions found in different cancers (FORBES *et al.* 2016; FORBES *et al.* 2017)(http://cancer.sanger.ac.uk/cosmic). The deletions have a median length of 1,255,650bps. For most of these deletions, the telomere proximal breakpoints map to a small chromosomal region (Chromosomal coordinates: Chr9:20000000-22000000) left of CDKN2A where the genes MTAP and MIR31HG are located. The right breakpoints spread over a larger region centromere proximal of CDKN2A. The fact that these deletions are not cancer specific suggests that something about the chromosomal structure in this region promotes these aberrations.

MATERIALS AND METHODS

Data analysis. For this analysis, we relied on the data presented on Catalogue of Somatic Mutations in Cancer (COSMIC) website (http://cancer.sanger.ac.uk/cosmic). For the data presented in Figure 1A we used the tool "Cancer Browser" then interrogated every cancer visually for the top 15 mutated genes. We recorded how many times mutation in a gene appeared in each cancer. For the data in Figure 1B we recorded all the samples that had any CDKN2A mutation (CNV, point mutation, etc.) from all sequenced cancer samples. Figure 1C shows the subset of the CDKN2A samples in Figure 1B that show only homozygous deletion.

Deletion breakpoints were downloaded from the COSMIC CONAN (Copy Number Analysis) database present on the website (https://cancer.sanger.ac.uk/cosmic/conan/search). All deletion breakpoints for CDKN2A were downloaded as an Excel file. For the data in Figure 2 the search term was "9:1-40000000" (Genomic Region). For the data in Figures 3, 4, 5 and 6 the search term was "CDKN2A" (HGNC Gene Symbol). All data were exported as CSV format. The data deposited on CONAN were derived from an Affymetrix SNP 6.0 Array and analyzed with PICNIC (GREENMAN *et al.* 2010) and ASCAT (VAN LOO *et al.* 2010).

The genomic DNA coordinates snapshot presented in Figure 5 was downloaded from UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgGateway) with the search prompt "chr9:18,000,000-32,500,000". The following tracts were on: Base position (dense), GENECODE V24 (pack), GTEx Gene (dense), and Segmental Dups (pack).

All data were analyzed and graphed in Excel.

Raw data availability. The data can be accessed on COSMIC CONAN or UCSC Genome Browser using the prompts described in the above Data Analysis section. The COSMIC files can be downloaded in *.csv* format. Each file contains 8 columns. The "# Sample" column lists the cancer subject identifier. The "Tissue" column lists the cancer tissues analyzed. The "Segment start" and "Segment end" columns list the chromosomal coordinates for the aberrations. These coordinates are at the resolution of the array. The "Total copy number" column lists the number of alleles identified. The "Minor allele" column represents the copies of the least frequent allele. The "Classification" column lists the type of aberration: HD=homozygous deletion, AMP= amplification, LOH= loss of heterozygosity.

RESULTS

A high percentage of chromosomal aberrations that include the CDKN2A locus are homozygous deletions. Interrogation of the COSMIC mutation database for the GRCh38 version of the human genome assembly revealed the top 15 mutated loci in 45 different cancer tissues (**Fig. 1A**). Of these 15 loci, the most prominent were TP53, which appeared in 36 tissues, and CDKN2A, which appeared in 18 tissues. COSMIC reports data for cancer genomes of several clinical samples from each tissue. When we examined the tissue distribution of CDKN2A aberrations independently, we found that the cancer samples with the highest percentage of CDKN2A aberrations were pleura (38%) and skin (18%) (**Fig. 1B**). The central nervous system (CNS) came in third with 12% of the sequenced samples having aberrations in the CDNK2A locus. However, the CNS had the most sequenced samples (6502 total samples). When we categorized the type of DNA damage responsible for the reported CDKN2A aberrations, we found that pleura showed 76% and skin showed 74% homozygous deletions (**Fig. 1C**). Furthermore, 50% of the CNS aberrations were large homozygous deletions; the remaining percentages representing non-deletion types of aberrations (e.g. point mutations).

A concentration of CNV breakpoints in a narrow region on Chromosome 9. To understand whether Chromosome 9 is more prone to breaking in certain regions, we first analyzed the distribution of the breakpoints in all cancers reported on COSMIC (**Fig. 2**). We used data deposited on the Copy Number Analysis (CONAN) database that catalogues only the copy number variations (CNVs) (FORER *et al.* 2010) acquired with an Affymetrix SNP6.0 array and reports CNV segment start and segment end for all breakpoints. When we generated scatter plots of both the left and right breakpoints of the *p* arm of Chromosome 9, we found that many concentrates in a narrow region between coordinates 20000000 and 25000000 (**Fig.2 A, B**). Note that this pattern is unique to this region and is not found anywhere on the *p* arm of Chromosome 9.

Mapping the position of breakpoints to the left and right of CDKN2A. The high incidence of homozygous deletions that span the CDNK2A locus as well as the concentration of breakpoints found here suggest that the chromosome is prone to breaking in this region. The Chromosome 9 p21.1 and p21.2 (**Fig.3A**) regions have been previously reported to be hotspots for large genomic aberrations (SASAKI *et al.* 2003; LEE *et al.* 2015). We next analyzed the CNVs that include only the CDKN2A locus. Out of a total of 1479 aberrations reported for the CDKN2A locus in all cancers tested, 20 (1.35%) are amplifications (AMP), 1397 (94.46%) are homozygous deletions (HD) and 62 (4.19%) are loss of heterozygosity (LOH). When we mapped all the reported CNVs for the CDNK2A locus, we found that, unlike LOH and AMP, HD aberrations localize to a narrow region of the *p* arm of Chromosome 9 with only two homozygous deletions extending past the 9p21.1 region (**Fig. 3B, C**). Furthermore, the types of aberrations are not tissue specific.

A close analysis of the homozygous deletions reveals that the break interval on the left side of CDKN2A (telomere proximal) maps to a much narrower region than the right (centromere proximal) break interval (**Fig. 3B**). Note that this is not a consequence of graphing method because in this figure each line shows position start and position end of the homozygous

deletion. Clearly, the left breakpoints are more clustered that the right breakpoints (**Supplementary Figure S1**).

Analysis of the chromosomal region spanning the left and right breakpoints. We next analyzed the chromosomal region where most of the breakpoints are found using a snapshot from the UCSC genome browser (**Fig. 4A**). The break interval where most left breakpoints occur includes the MTAP, IFN transcripts, and MIR31HG while the right breakpoints include ELAVL2, and TEK genes. When we graphed only the left break point interval for all reported CDKN2A homozygous deletions, we found that it clusters within 2.5Mbs (chromosomal coordinates 19400000-22000000) (**Fig. 4B**). The right breakpoint spans over a region (Ch9: 21977195-32942228, approx. 10965033bp) which is about four times larger than the region spanning the left breakpoint (**Fig.4 A, B**). Figure 4 shows the breakpoint density as number of breaks per genomic base pair. Clearly most of the left breakpoints are concentrated in the MTAP region.

We next calculated the percentage of breakpoints in each region. Most left break points (1098/1398, 78.5%) occur approximately between coordinates 20995956-21937651, which is just under 1Mb (**Fig. 4B**). A smaller percentage of breakpoints (204/1398, 14.6%) occur in the region of the FOCAD, MLLT3, and SLC24A2 loci. Few breakpoints (96/1398, 6.9%) occur between CDKN2A and the end of MTAP. Within the right region we find that 70.7% of the breakpoints occur between CDKN2A and ELAVL2, 1.6% occur in the ELAVL2 region, 15.4% occur between ELAVL2 and the TEK region and 1.36% occur in the TEK region. The rest of the breakpoints are to the right of the TEK region.

We next analyzed the nature of the repetitive elements in this region. We did not see an obvious concentration of SINE, LINE and LTR elements (data now shown). However, we did see a high concentration of segmental duplications that localize within the IFNA transcripts region, where about 25% of the left breakpoints are found and within the TEK region where only 1.36% of right breakpoints are found (**Fig.4A**). Segmental duplications have been proposed to arise from repair of damaged DNA replication forks (PAYEN *et al.* 2008; COSTANTINO *et al.* 2014). A correlation between the other approximately 74% of the breakpoints and segmental duplications cannot be made.

Correlation of chromosomal coordinates and deletion size. We next tried to understand whether there is any relationship between the size of the homozygous deletion and the position of the breakpoints. A correlation between the right and the left breakpoints showed that most breakpoints concentrate just to the left and right of CDKN2A (**Fig.5A**). Indeed, when we graphed the deletions by size, we did find that most of them are short (approximately 50%) with the breakpoints just to the left and right of CDKN2A (**Fig. 5B**). Remarkably, the short homozygous deletions have their left break point in the MTAP region telomere proximal of CDKN2A. Furthermore, we find that the homozygous deletion size increase is determined primarily by the position of the right breakpoint (**Fig.6**). Note that the left breakpoints remain concentrated in a small region while the right breakpoints move further and further to the right and this correlates with the increase in deletion size.

DISCUSSION

Here we provide evidence that the left breakpoints of homozygous deletions that include the CDKN2A locus are enriched in a region telomere proximal of the CDKN2A locus. Previous analyses have identified such aberrations in distinct tumors but here we present a more comprehensive map of these aberrations in a multitude of cancers. We show that these aberrations are not cancer specific but are probably related to the structure or other molecular transactions in the chromosomal region in which they occur.

Inactivation of the MLL (Mixed Lineage Leukemia) and related genes have been identified in different forms of leukemias (WINTERS AND BERNT 2017). Many of these genes are inactivated by various translocations which produces chimeric mRNAs. The most famous is the Philadelphia chromosome translocation which produces a fusion between ABL1 and BCR1 (KANG *et al.* 2016). Further analysis of translocation in leukemia cancers led to identification of a plethora of other genes which were named after the point of translocation. This includes MLLT3 (Multiple Lineage Leukemia Translocated to Ch. 3) (IIDA *et al.* 1993). MLLT3 characterized by tri-nucleotide repeats which may facilitate these forms of Translocations (WALKER *et al.* 1994). Several non-reciprocal translocations between MLLT3 and other chromosome loci have been identified (MEYER *et al.* 2013). The presence of an MLLT3 gene in the vicinity of CDKN2A may explain why some of the breakpoints localize in this region but does not explain most of the breakpoints which occur in the MTAP region or between MTAP and MIR31HG. Remarkably, MTAP is the only transcript that collides with CDKN2A which raises the possibility that collisions between transcriptional machineries may lead to a higher incidence of breaks in this region. However, in this analysis we do not have any data for this conclusion.

MIR31HG encodes a long-non-coding RNA with oncogenic properties that represses expression of p16(MONTES *et al.* 2015). MIR31HG dysregulation has been identified in many cancers including pancreatic (MONTES *et al.* 2015; NIE *et al.* 2016; YANG *et al.* 2016). Thus, it appears that the homozygous deletion events seen here simultaneously inactivate both CDKN2A and its regulators. MTAP encodes the enzyme methylthioadenosine phosphorylase which is required early in the purine biosynthesis pathway (BERTINO *et al.* 2011). Deletion of MTAP has been identified in many forms of cancers and is usually co-deleted with the CDKN2A locus (CHEN *et al.* 1996).

Several molecular mechanisms have been identified that can lead to homozygous deletions. One of the most preeminent mechanism reported in human cells to cause deletions is Microhomology Mediated Break Induced Replication (MMBIR) (HASTINGS *et al.* 2009a; VERDIN *et al.* 2013). MMBIR has also been proposed to be the major cause of copy number variations (HASTINGS *et al.* 2009a; ZHANG *et al.* 2009a; ZHANG *et al.* 2009b). What is attractive about MMBIR is that it does not require two ended breaks, and potential breaks resulting from replication fork stalling or collapsing usually produces one ended breaks. Because these data were generated with microarrays, breakpoints are only at the precision of the arrays which is at best within 15bps (SUDMANT *et al.* 2015). Therefore, this data cannot be analyzed for microhomologies.

Canonical break induced replication (BIR) requires more extensive homologous regions in the vicinity of 300 nucleotides (MEHTA AND HABER 2014). We visually inspected all breakpoints for longer tracts of homology, but we did not see large regions of homology characteristic of repair by break induced replication, or other forms of homologous recombination such as single strand annealing (MEHTA AND HABER 2014; BHARGAVA *et al.* 2016). Non-homologous end joining (NHEJ), a form of repair that is very efficient in human cells, is another mechanism that can cause copy number variations (HASTINGS *et al.* 2009b). A recently described Trans-Flip mechanism has been proposed to generate mainly short homozygous deletions (NORRIS *et al.* 2015b). We do find that the CDKN2A homozygous deletions are short and they fall in the range of the those proposed to occur by TransFlip so it is possible that this mechanism may contribute to these aberrations.

Here we provide evidence that the left breakpoints of homozygous deletions that include the CDKN2A locus are enriched in a region to the left of CDKN2A where MTAP is found. Human genomes are characterized by numerous other forms of chromosomal aberrations many of which include duplications and deletions. Deletion hotspots as the one described here are most likely also responsible for the copy number variations that accumulated in the human genome throughout evolution.

FIGURES

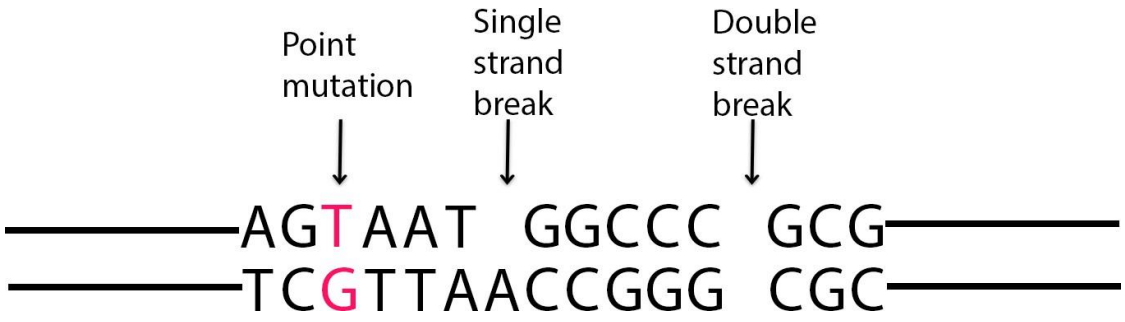


Figure 1. Major types of DNA damage. Three major types of DNA damage can be encountered in a double stranded chromosome. In point mutation the base on one strand has changed such that it violates the rules of base pairing A-T and G-C. In this case C has changed to T. In a single strand break, one of the two strands breaks. This can be easily repaired because the information is preserved on the other strand. In this case a *T* would be inserted opposite *A*. A double strand break constitutes severing of both strands. In this diagram the right part of the chromosome could be lost because it completely dissociates from the left part.

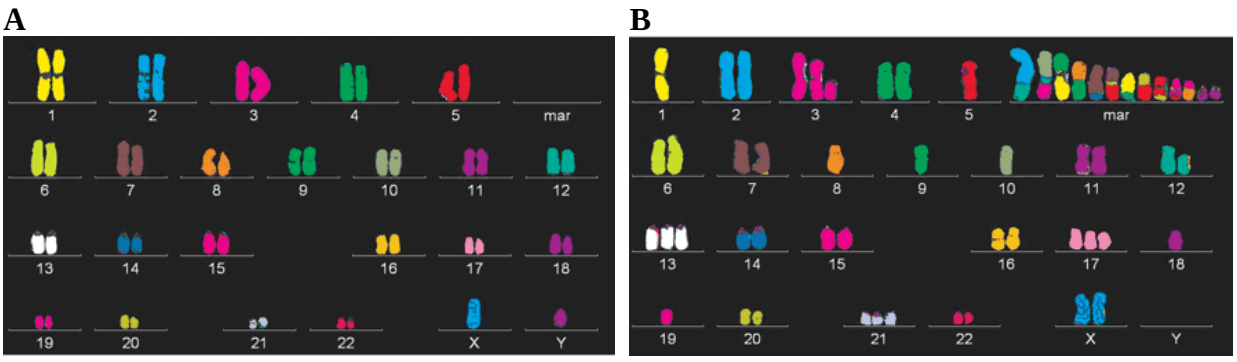


Figure 2. Microscopic chromosome karyotypes (Source: Biology Learning Center). (A) Each of the 22 chromosome pairs and the sex XY chromosomes are painted with a different color to be distinguished from each other under the microscope. (B) A cancer cell has reorganized its chromosomes such that many can no longer be classified as a normal 1 to 22 chromosome. It is said that this cancer cell has been transformed.

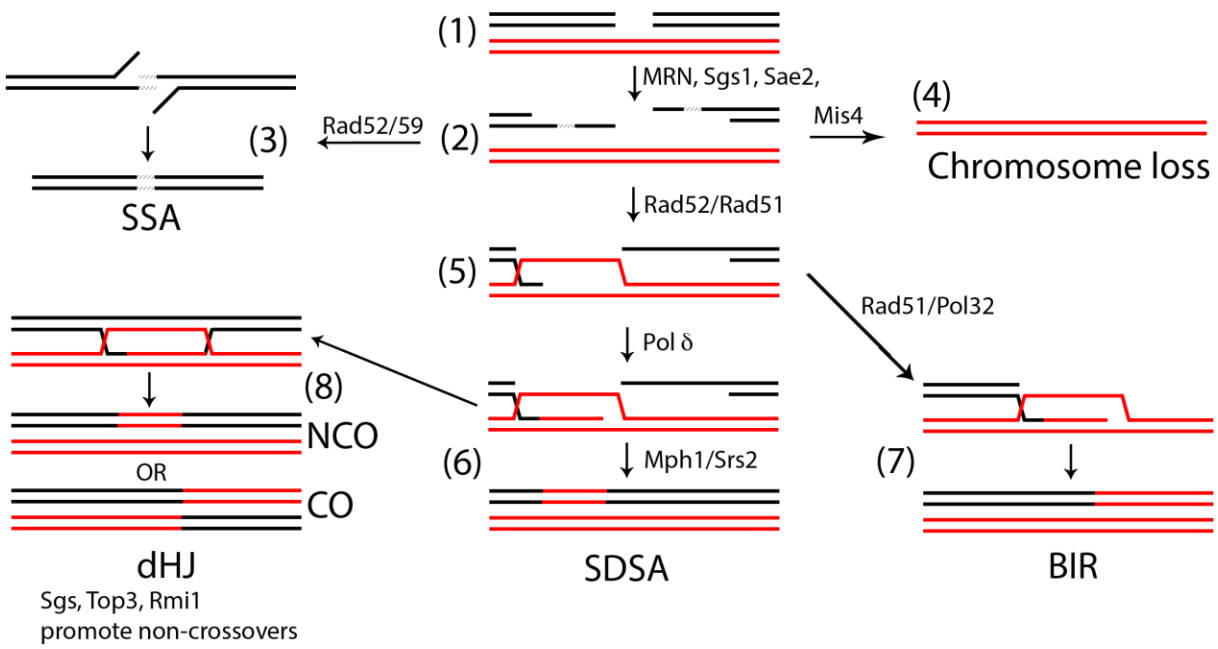


Figure 3. Pathways of repair of DNA double strand breaks. A diploid cell with two homologous chromosomes, black and red, sustains a double strand break (DSB) in the black chromosome (1). The DSB is first resected to expose ssDNA required for invasion of donor regions (2). If direct repeats (shaded areas) exist on the same chromosome, the break may be repaired by single strand annealing (SSA) (3). If repair fails, the chromosome may be lost (4). When homology is found elsewhere or on the other homologue (red), the broken ends may invade this region (the donor sequence) (5). In synthesis-dependent strand annealing (SDSA) (6) the invading strand may copy a small region then release and re-anneal. In break-induced replication (BIR) (7) the invading strand may copy to the end of the red chromosome. In this case the right portion of the broken black chromosome is lost. Occasionally a more complex double Holiday Junction (dJH) may be established (8), the resolution of which can result in crossovers (CO) or non-crossovers (NCO). Note that some of these repair outcomes may lead to loss of heterozygosity meaning that the black sequence has been converted to red. If the red sequence contains a recessive non-functional allele, some of these outcomes will convert the functional black allele to the non-functional red allele resulting in complete inactivation of the gene. Some of the genetic requirements for each pathway are indicated.

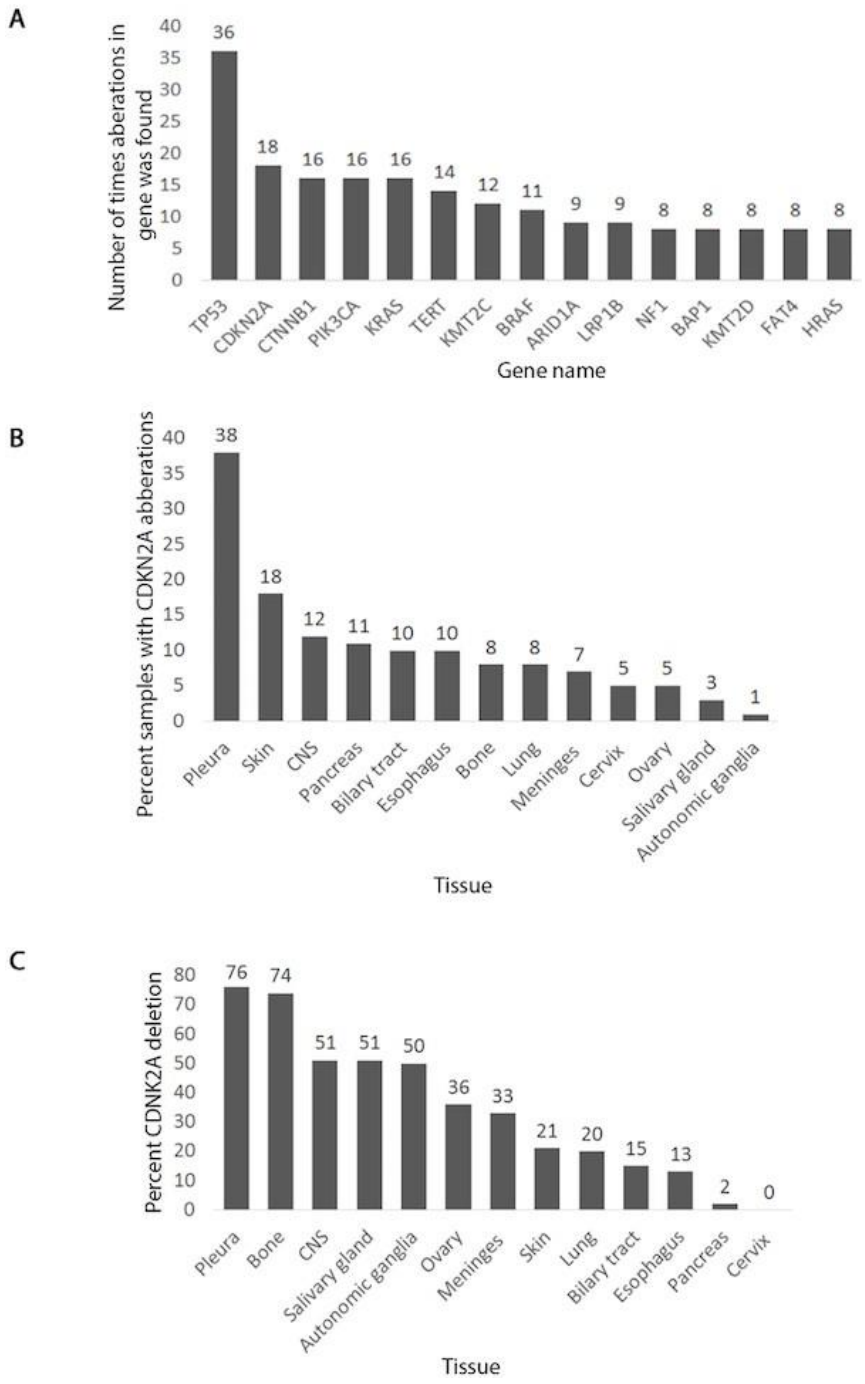


Figure 4. The CDKN2A locus is mutated in many cancers. (A) The top 15 mutations that appeared in 45 different cancers as reported by the COSMIC database. (B) Percent cancer samples that have any CDKN2A aberration out of all sequenced samples in each tissue. Only the tissues that have over 1% CDKN2A aberrations are shown. (C) Percent samples with CDKN2A homozygous deletions of all the samples with CDKN2A aberrations in (B)

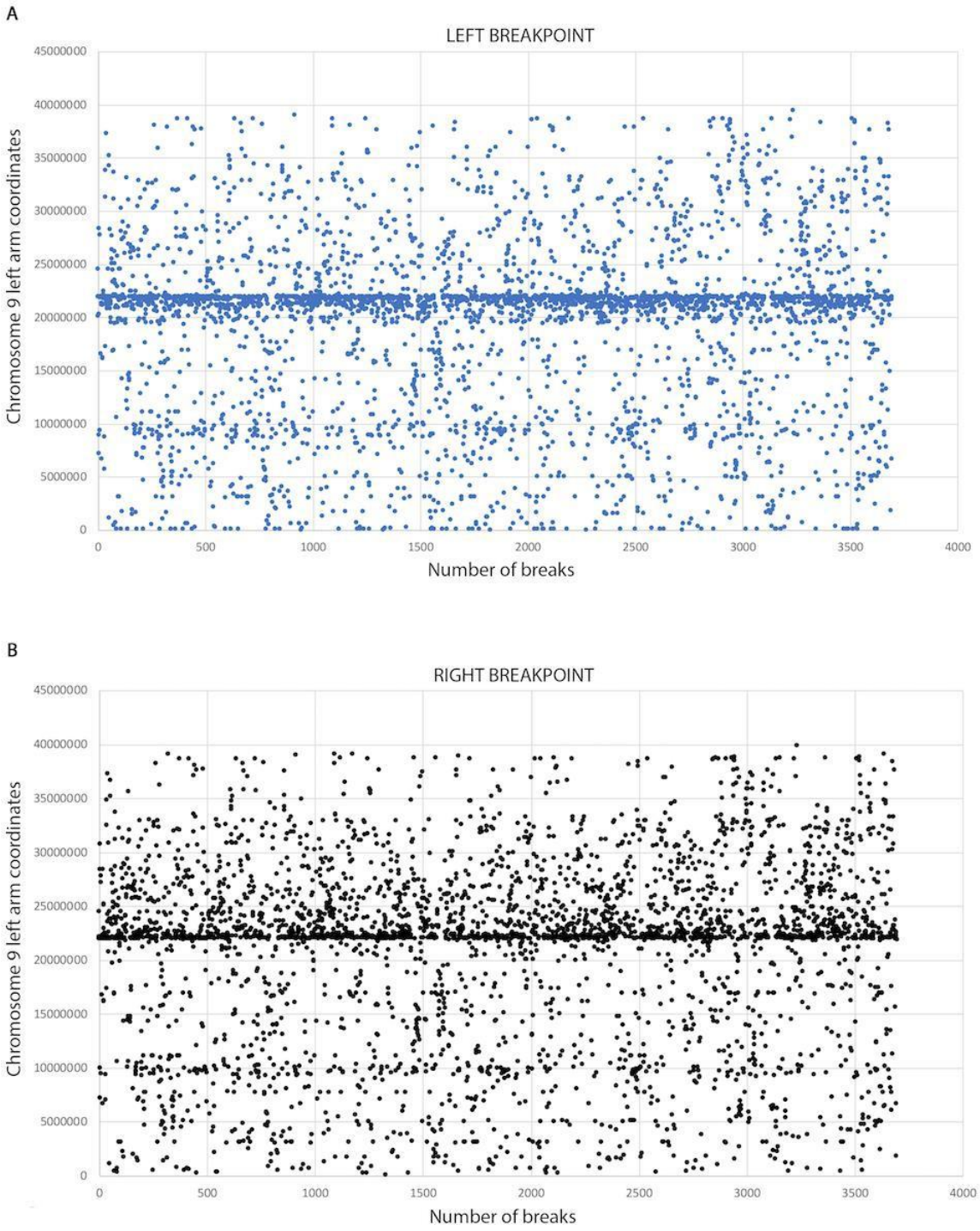


Figure 5. Many CNV breakpoints concentrate in a narrow range on Chromosome 9. Scatter graphs of all CONAN reported CNV breakpoints on the *p* arm of Chromosome 9. (A) Left breakpoint (proximal to telomere of p arm) of CNVs. Note that there is a concentration of breakpoints between coordinates 20,000,000 and 25,000,000. (B) Right breakpoint (centromere proximal) of CNVs.

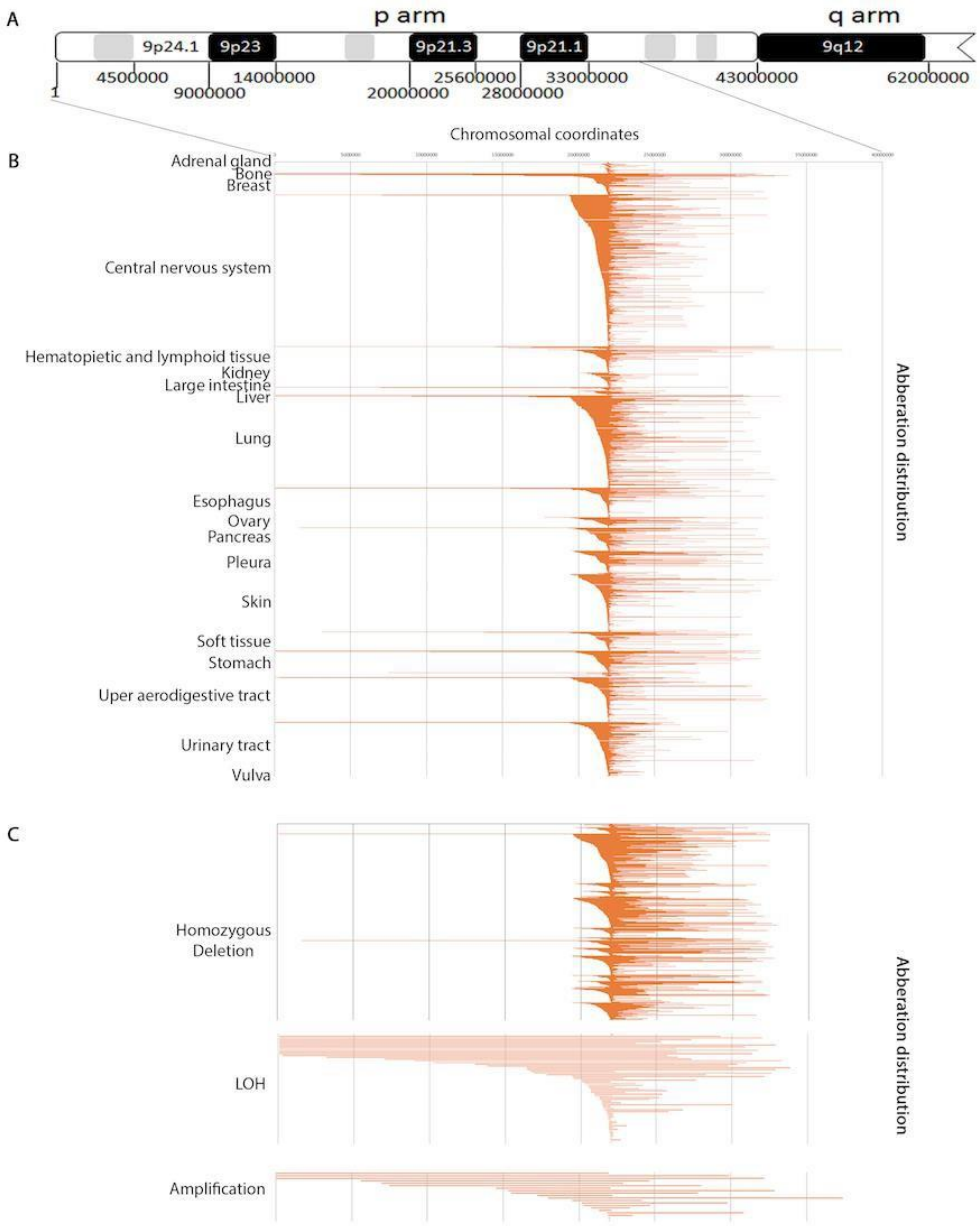


Figure 6. Chromosomal aberrations that include the CDKN2A locus. (A) General map of the chromosome 9 p arm. Chromosomal coordinates are shown in base pairs. (B) A map of all the chromosomal aberrations reported on CONAN by tissue type. Included are loss of heterozygosity (LOH), homozygous deletion (HD) and amplification (AMP). The lines represent the aberration between the left and right breakpoints as reported on CONAN. (C) Maps of aberrations in B separated by aberration type.

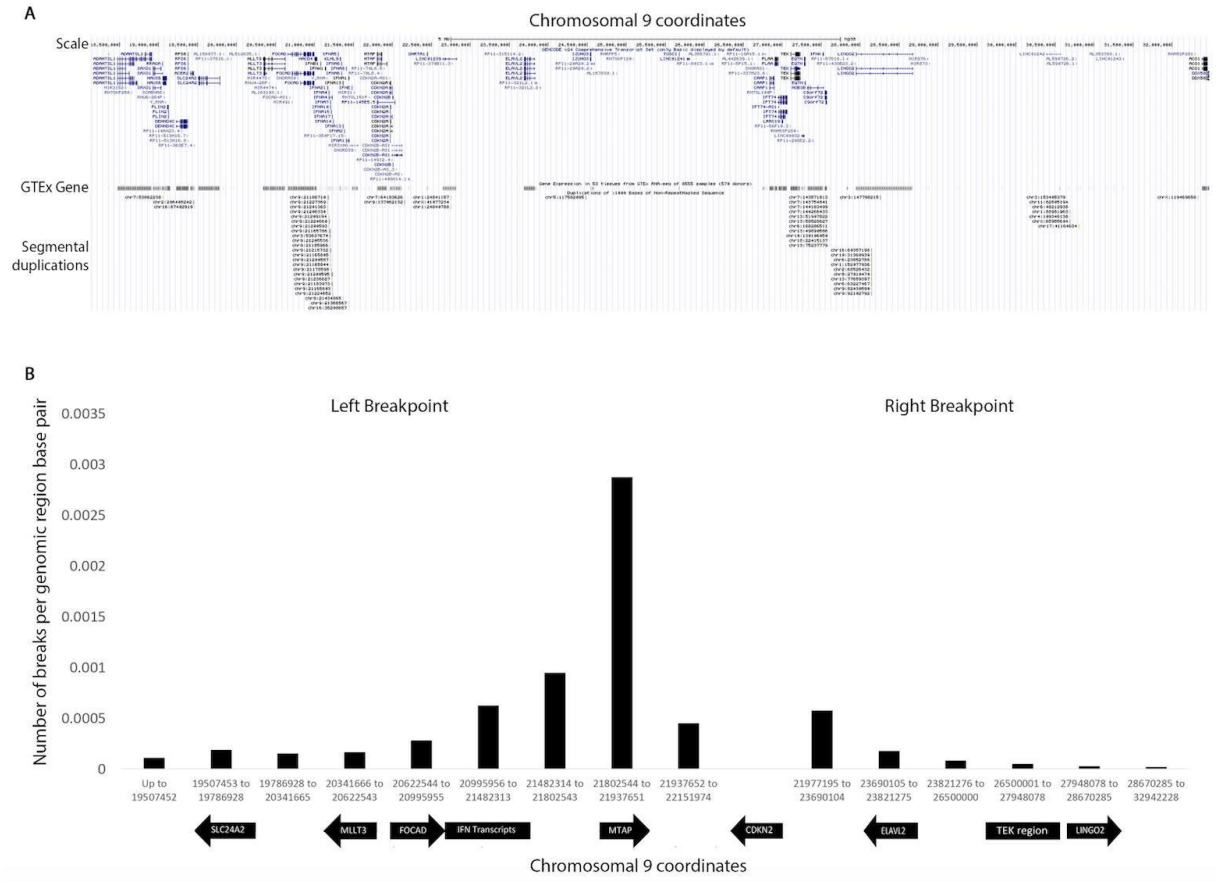


Figure 7. Distribution of breakpoints of CNVs that include CDKN2A over the chromosomal region. (A) UCSC genome browser snapshot of the chromosomal regions neighboring CDKN2A with a window between coordinates 9:18,000,000-32,500,000. Showing are the base positions, coding genes, GTEx Gene expression and segmental duplications. (B) Density of breakpoint distribution to the left and right of CDKN2A. Because this graph is not superimposable on the genome snapshot in A some of the genomic loci are shown on the X-axis.

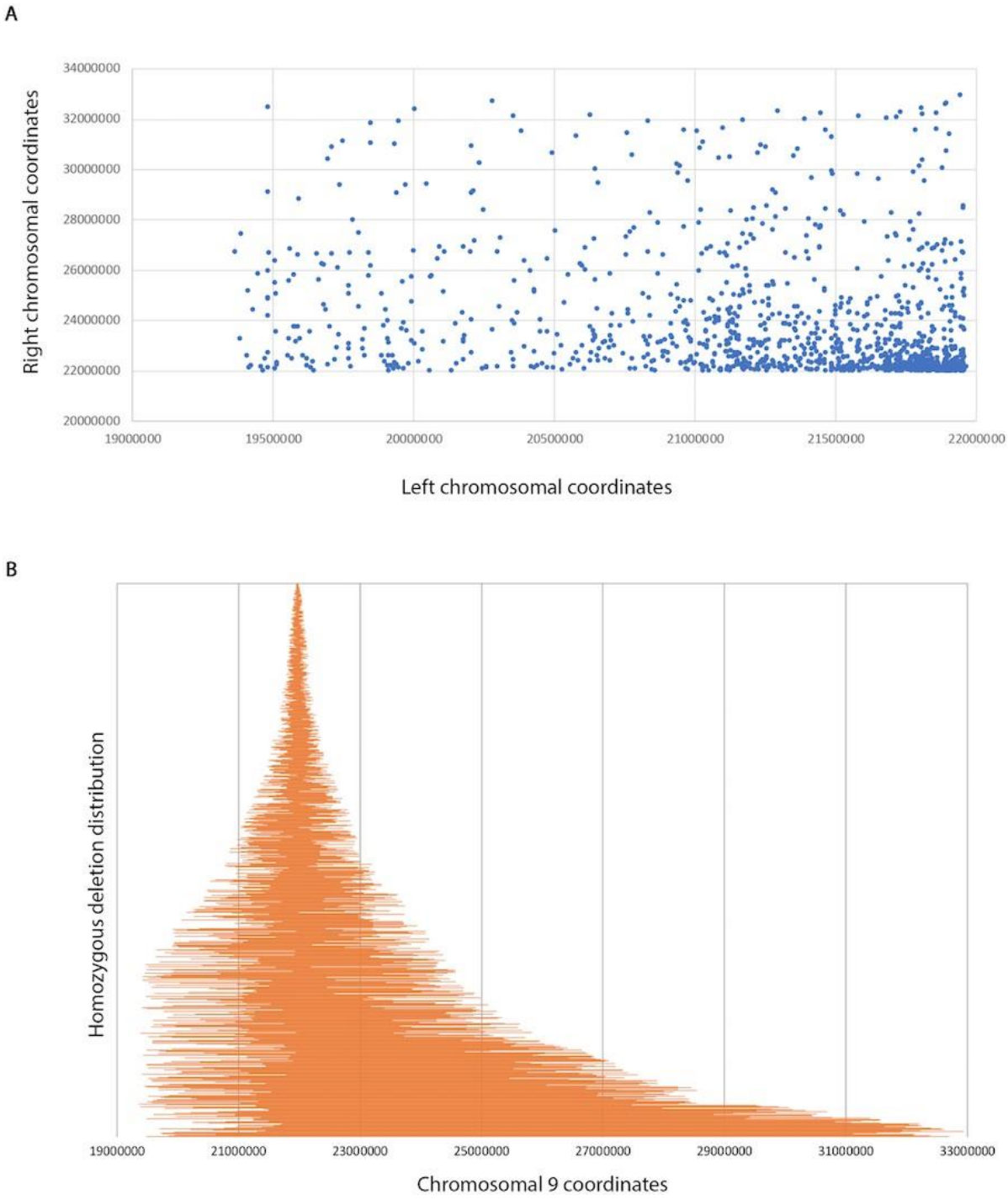


Figure 8. Correlation between the left breakpoint and the right breakpoint. (A) X-Y scatter graph of the left and right breakpoint for all homozygous deletions that include CDKN2A. X-axis shows the chromosomal coordinates left of CDKN2A and Y-axis the coordinates right of CDKN2A. (B) A graph of the homozygous deletions by size. Note that approximately 50% of the deletions are short and the right breakpoints spread over a larger region.

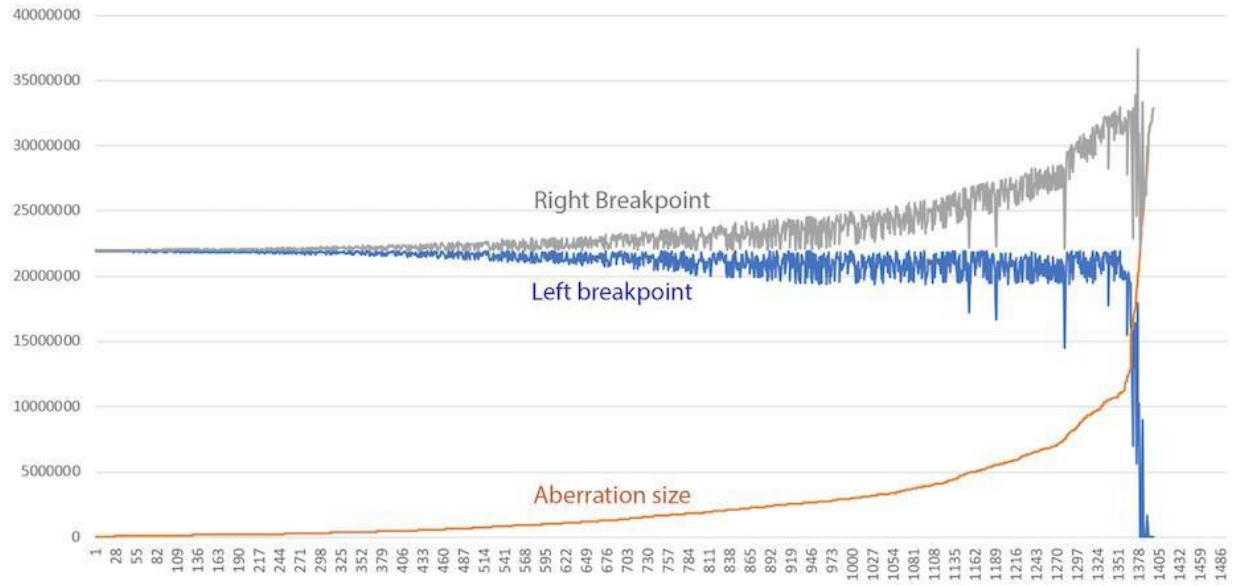
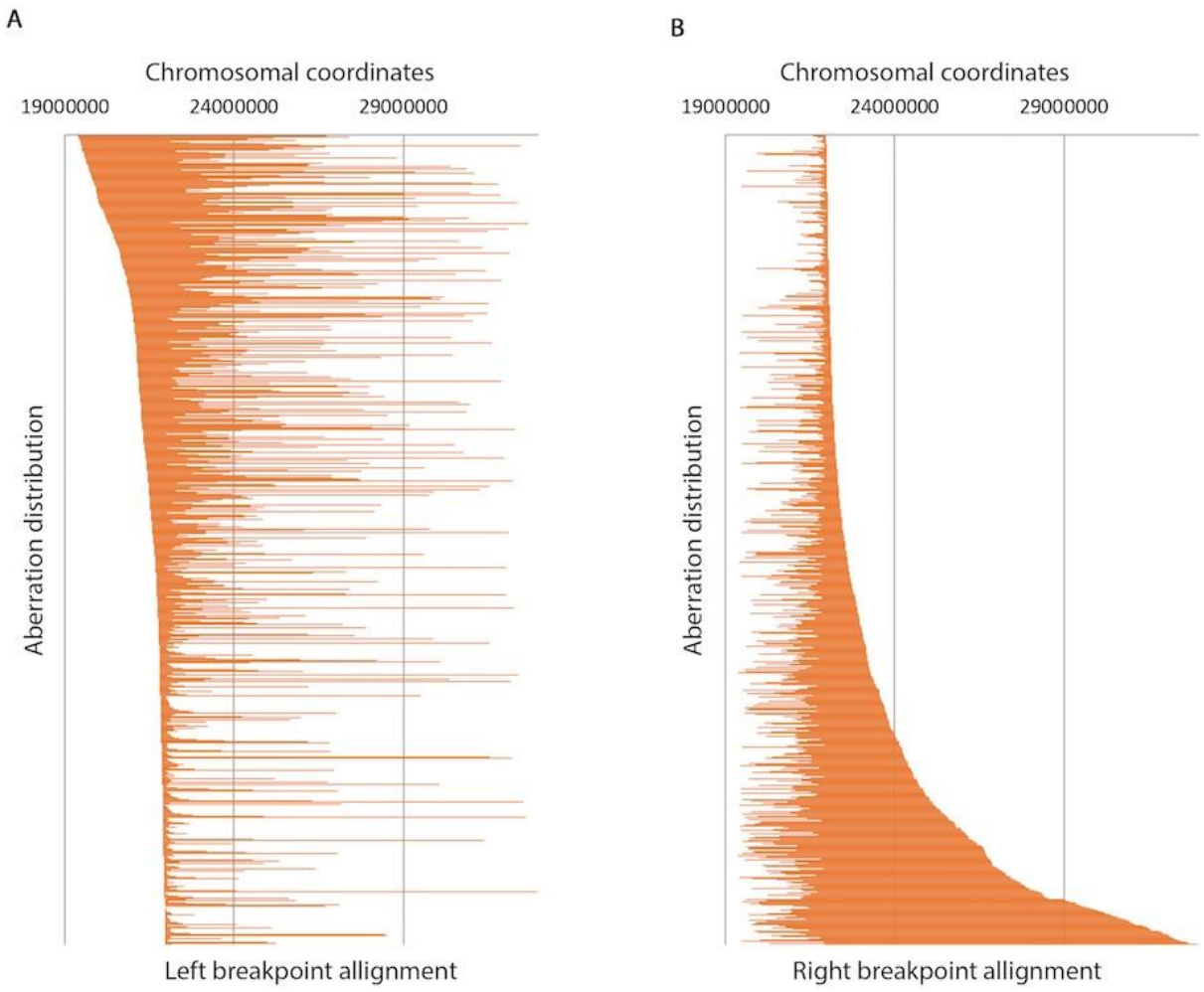


Figure 9. Correlation of breakpoints and homozygous deletion size. Graph showing correlation between length of homozygous deletion (orange) its left breakpoint (blue) and its right breakpoint (grey). Note that the left breakpoints are clustered, and the deletions size increase is dependent on the position of the right breakpoint.



Supplementary Figure S1. The left breakpoints of homozygous deletions are concentrated to a narrow region than the right breakpoints. (A) Graph showing the distribution of homozygous deletions organized by the position of the left breakpoints. (B) Graph showing the distribution of homozygous deletions organized by the position of the right breakpoints.

REFERENCES

Bertino, J. R., W. R. Waud, W. B. Parker and M. Lubin, 2011 Targeting tumors that lack methylthioadenosine phosphorylase (MTAP) activity Current strategies. Cancer Biology & Therapy 11**:** 627-632.

Bhargava, R., D. O. Onyango and J. M. Stark, 2016 Regulation of Single-Strand Annealing and its Role in Genome Maintenance. Trends Genet 32**:** 566-575.

Bhattacharjee, S., and S. Nandi, 2016 Choices have consequences: the nexus between DNA repair pathways and genomic instability in cancer. Clin Transl Med 5**:** 45.

Bieging, K. T., S. S. Mello and L. D. Attardi, 2014 Unravelling mechanisms of p53-mediated tumour suppression. Nat Rev Cancer 14**:** 359-370.

Chang, H. H. Y., N. R. Pannunzio, N. Adachi and M. R. Lieber, 2017 Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Biol 18**:** 495-506.

Chapman, J. R., M. R. Taylor and S. J. Boulton, 2012 Playing the end game: DNA double-strand break repair pathway choice. Mol Cell 47**:** 497-510.

Chen, Z. H., H. Y. Zhang and T. M. Savarese, 1996 Gene deletion chemoselectivity: Codeletion of the genes for p16(INK4) methylthioadenosine phosphorylase, and the alpha- and beta-interferons in human pancreatic cell carcinoma lines and its implications for chemotherapy. Cancer Research 56**:** 1083-1090.

Cicenas, J., K. Kvederaviciute, I. Meskinyte, E. Meskinyte-Kausiliene, A. Skeberdyte *et al.*, 2017 KRAS, TP53, CDKN2A, SMAD4, BRCA1, and BRCA2 Mutations in Pancreatic Cancer. Cancers (Basel) 9.

Conrad, D. F., C. Bird, B. Blackburne, S. Lindsay, L. Mamanova *et al.*, 2010 Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. Nat Genet 42**:** 385-391.

Costantino, L., S. K. Sotiriou, J. K. Rantala, S. Magin, E. Mladenov *et al.*, 2014 Break-induced replication repair of damaged forks induces genomic duplications in human cells. Science 343**:** 88-91.

Forbes, S. A., D. Beare, N. Bindal, S. Bamford, S. Ward *et al.*, 2016 COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. Curr Protoc Hum Genet 91**:** 10 11 11-10 11 37.

Forbes, S. A., D. Beare, H. Boutselakis, S. Bamford, N. Bindal *et al.*, 2017 COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res 45**:** D777-D783.

Forer, L., S. Schonherr, H. Weissensteiner, F. Haider, T. Kluckner *et al.*, 2010 CONAN: copy number variation analysis software for genome-wide association studies. Bmc Bioinformatics 11.

Friedberg, E. C., 2008 A brief history of the DNA repair field. Cell Res 18**:** 3-7.

Gast, A., D. Scherer, B. Chen, S. Bloethner, S. Melchert *et al.*, 2010 Somatic alterations in the melanoma genome: a high-resolution array-based comparative genomic hybridization study. Genes Chromosomes Cancer 49**:** 733-745.

Greenman, C. D., G. Bignell, A. Butler, S. Edkins, J. Hinton *et al.*, 2010 PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics 11**:** 164-175.

Hastings, P. J., G. Ira and J. R. Lupski, 2009a A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genet 5**:** e1000327.

Hastings, P. J., J. R. Lupski, S. M. Rosenberg and G. Ira, 2009b Mechanisms of change in gene copy number. Nat Rev Genet 10**:** 551-564.
Hromas, R., E. Williamson, S. H. Lee and J. Nickoloff, 2016 Preventing the Chromosomal Translocations That Cause Cancer. Trans Am Clin Climatol Assoc 127**:** 176-195.
Iida, S., M. Seto, K. Yamamoto, H. Komatsu, A. Tojo *et al.*, 1993 Mllt3 Gene on 9p22 Involved in T(9-11) Leukemia Encodes a Serine Proline-Rich Protein Homologous to Mllt1 on 19p13. Oncogene 8**:** 3085-3092.
Jackson, S. P., and J. Bartek, 2009 The DNA-damage response in human biology and disease. Nature 461**:** 1071-1078.
Kang, Z. J., Y. F. Liu, L. Z. Xu, Z. J. Long, D. Huang *et al.*, 2016 The Philadelphia chromosome in leukemogenesis. Chinese Journal of Cancer 35.
Kashiwagi, H., and K. Uchida, 2000 Genome-wide profiling of gene amplification and deletion in cancer. Hum Cell 13**:** 135-141.
Kennedy, S. R., L. A. Loeb and A. J. Herr, 2012 Somatic mutations in aging, cancer and neurodegeneration. Mech Ageing Dev 133**:** 118-126.
Kim, W. Y., and N. E. Sharpless, 2006 The regulation of INK4/ARF in cancer and aging. Cell 127**:** 265-275.
Kowalczykowski, S. C., 2015 An Overview of the Molecular Mechanisms of Recombinational DNA Repair. Cold Spring Harb Perspect Biol 7.
Lee, B., K. Yoon, S. Lee, J. M. Kang, J. Kim *et al.*, 2015 Homozygous deletions at 3p22, 5p14, 6q15, and 9p21 result in aberrant expression of tumor suppressor genes in gastric cancer. Genes Chromosomes Cancer 54**:** 142-155.
Li, P. C., R. C. Petreaca, A. Jensen, J. P. Yuan, M. D. Green *et al.*, 2013 Replication fork stability is essential for the maintenance of centromere integrity in the absence of heterochromatin. Cell Rep 3**:** 638-645.
Lindahl, T., and D. E. Barnes, 2000 Repair of endogenous DNA damage. Cold Spring Harb Symp Quant Biol 65**:** 127-133.
Lukas, J., and J. Bartek, 2009 DNA repair: New tales of an old tail. Nature 458**:** 581-583.
Lupski, J. R., 1998 Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. Trends Genet 14**:** 417-422.
Malumbres, M., and M. Barbacid, 2005 Mammalian cyclin-dependent kinases. Trends Biochem Sci 30**:** 630-641.
Malumbres, M., and M. Barbacid, 2009 Cell cycle, CDKs and cancer: a changing paradigm. Nat Rev Cancer 9**:** 153-166.
Malumbres, M., E. Harlow, T. Hunt, T. Hunter, J. M. Lahti *et al.*, 2009 Cyclin-dependent kinases: a family portrait. Nat Cell Biol 11**:** 1275-1276.
Mehta, A., and J. E. Haber, 2014 Sources of DNA double-strand breaks and models of recombinational DNA repair. Cold Spring Harb Perspect Biol 6**:** a016428.
Meyer, C., J. Hofmann, T. Burmeister, D. Groger, T. S. Park *et al.*, 2013 The MLL recombinome of acute leukemias in 2013. Leukemia 27**:** 2165-2176.
Montes, M., M. M. Nielsen, G. Maglieri, A. Jacobsen, J. Hojfeldt *et al.*, 2015 The lncRNA MIR31HG regulates p16(INK4A) expression to modulate senescence. Nature Communications 6.
Mountzios, G., T. Rampias and A. Psyrri, 2014 The mutational spectrum of squamous-cell carcinoma of the head and neck: targetable genetic events and clinical impact. Ann Oncol 25**:** 1889-1900.

Nie, F. Q., S. J. Ma, M. Xie, Y. W. Liu, W. De *et al.*, 2016 Decreased long noncoding RNA MIR31HG is correlated with poor prognosis and contributes to cell proliferation in gastric cancer. Tumor Biology 37**:** 7693-7701.

Norris, A. L., H. Kamiyama, A. Makohon-Moore, A. Pallavajjala, L. A. Morsberger *et al.*, 2015a Transflip mutations produce deletions in pancreatic cancer. Genes Chromosomes Cancer 54**:** 472-481.

Norris, A. L., H. Kamiyama, A. Makohon-Moore, A. Pallavajjala, L. A. Morsberger *et al.*, 2015b Transflip mutations produce deletions in pancreatic cancer. Genes Chromosomes Cancer.

Ortega, S., M. Malumbres and M. Barbacid, 2002 Cyclin D-dependent kinases, INK4 inhibitors and cancer. Biochim Biophys Acta 1602**:** 73-87.

Orthwein, A., S. M. Noordermeer, M. D. Wilson, S. Landry, R. I. Enchev *et al.*, 2015 A mechanism for the suppression of homologous recombination in G1 cells. Nature 528**:** 422-426.

Pastink, A., J. C. Eeken and P. H. Lohman, 2001 Genomic integrity and the repair of double-strand DNA breaks. Mutat Res 480-481**:** 37-50.

Payen, C., R. Koszul, B. Dujon and G. Fischer, 2008 Segmental duplications arise from Pol32-dependent repair of broken forks through two alternative replication-based mechanisms. PLoS Genet 4**:** e1000175.

Pines, J., 1994a The cell cycle kinases. Semin Cancer Biol 5**:** 305-313.

Pines, J., 1994b Protein kinases and cell cycle control. Semin Cell Biol 5**:** 399-408.

Pines, J., and T. Hunter, 1994 The differential localization of human cyclins A and B is due to a cytoplasmic retention signal in cyclin B. EMBO J 13**:** 3772-3781.

Poi, M. J., T. J. Knobloch, M. T. Sears, B. M. Warner, L. K. Uhrig *et al.*, 2015 Alterations in RD(INK4/ARF) -mediated en bloc regulation of the INK4-ARF locus in human squamous cell carcinoma of the head and neck. Mol Carcinog 54**:** 532-542.

Ranjha, L., S. M. Howard and P. Cejka, 2018 Main steps in DNA double-strand break repair: an introduction to homologous recombination and related processes. Chromosoma 127**:** 187-214.

Sasaki, S., Y. Kitagawa, Y. Sekido, J. D. Minna, H. Kuwano *et al.*, 2003 Molecular processes of chromosome 9p21 deletions in human cancers. Oncogene 22**:** 3792-3798.

Shibata, A., 2017 Regulation of repair pathway choice at two-ended DNA double-strand breaks. Mutat Res 803-805**:** 51-55.

Stransky, N., A. M. Egloff, A. D. Tward, A. D. Kostic, K. Cibulskis *et al.*, 2011 The mutational landscape of head and neck squamous cell carcinoma. Science 333**:** 1157-1160.

Sudmant, P. H., T. Rausch, E. J. Gardner, R. E. Handsaker, A. Abyzov *et al.*, 2015 An integrated map of structural variation in 2,504 human genomes. Nature 526**:** 75-81.

Symington, L. S., and J. Gautier, 2011 Double-strand break end resection and repair pathway choice. Annu Rev Genet 45**:** 247-271.

Tubbs, A., and A. Nussenzweig, 2017 Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell 168**:** 644-656.

Van Loo, P., S. H. Nordgard, O. C. Lingjaerde, H. G. Russnes, I. H. Rye *et al.*, 2010 Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A 107**:** 16910-16915.

Verdin, H., B. D'Haene, D. Beysen, Y. Novikova, B. Menten *et al.*, 2013 Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. PLoS Genet 9**:** e1003358.

Visconti, R., R. Della Monica and D. Grieco, 2016 Cell cycle checkpoint in cancer: a therapeutically targetable double-edged sword. J Exp Clin Cancer Res 35**:** 153.
Walker, G. J., M. K. Walters, J. M. Palmer and N. K. Hayward, 1994 The Mllt3 Gene Maps between D9s156 and D9s171 and Contains an Unstable Polymorphic Trinucleotide Repeat. Genomics 20**:** 490-491.
Wang, G., and K. M. Vasquez, 2017 Effects of Replication and Transcription on DNA Structure-Related Genetic Instability. Genes (Basel) 8.
Wang, W., J. M. Daley, Y. Kwon, X. Xue, D. S. Krasner *et al.*, 2018 A DNA nick at Ku-blocked double-strand break ends serves as an entry site for exonuclease 1 (Exo1) or Sgs1-Dna2 in long-range DNA end resection. J Biol Chem 293**:** 17061-17069.
Ward, J. F., 1988 DNA damage produced by ionizing radiation in mammalian cells: identities, mechanisms of formation, and reparability. Prog Nucleic Acid Res Mol Biol 35**:** 95-125.
Winters, A. C., and K. M. Bernt, 2017 MLL-Rearranged Leukemias-An Update on Science and Clinical Approaches. Frontiers in Pediatrics 5.
Xie, H., P. S. Rachakonda, B. Heidenreich, E. Nagore, A. Sucker *et al.*, 2016 Mapping of deletion breakpoints at the CDKN2A locus in melanoma: detection of MTAP-ANRIL fusion transcripts. Oncotarget 7**:** 16490-16504.
Yang, H., P. Liu, J. Zhang, X. Peng, Z. Lu *et al.*, 2016 Long noncoding RNA MIR31HG exhibits oncogenic property in pancreatic ductal adenocarcinoma and is negatively regulated by miR-193b. Oncogene 35**:** 3647-3657.
Zhang, F., C. M. Carvalho and J. R. Lupski, 2009a Complex human chromosomal and genomic rearrangements. Trends Genet 25**:** 298-307.
Zhang, F., M. Khajavi, A. M. Connolly, C. F. Towne, S. D. Batish *et al.*, 2009b The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. Nat Genet 41**:** 849-853.
Biology Learning Center (http://blc.arizona.edu/)